On constraining estimates of climate sensitivity with present-day observations through model weighting

Daniel Klocke [*]

Max Planck Institute for Meteorology, Hamburg, Germany

Robert Pincus

Cooperative Institute for Research in Environmental Sciences, University of Colorado and NOAA Earth System Research Laboratory, Physical Sciences Division, Boulder, CO

Johannes Quaas

Max Planck Institute for Meteorology, Hamburg, Germany

[*] *Corresponding Author Address:* Daniel Klocke, Max Planck Institute for Meteorology, Bundesstr. 53, 20146 Hamburg, Germany. email: Daniel.Klocke@zmaw.de

ABSTRACT

The distribution of model-based estimates of equilibrium climate sensitivity has not changed substantially in more than 30 years. Efforts to narrow this distribution by weighting projections according to measures of model fidelity have so far failed, largely because climate sensitivity is independent of current measures of skill in current ensembles of models. Here we provide a cautionary example showing that measures of model fidelity that are effective at narrowing the distribution of future projections (because they are systematically related to climate sensitivity in an ensemble of models) may be poor measures of the likelihood that a model will provide an accurate estimate of climate sensitivity (and so degrade distributions of projections if they are used as weights). Furthermore, it appears unlikely that statistical tests alone can identify robust measures of likelihood. We consider two ensembles: one obtained by perturbing parameters in a single climate model, and a second containing the majority of the world's climate models. The simple ensemble reproduces many aspects of the multi-model ensemble, including the distributions of skill in reproducing the present-day climatology of clouds and radiation, the distribution of climate sensitivity, and the dependence of climate sensitivity on certain cloud regimes. By restricting error measures to those regimes we can identify tighter relationships between climate sensitivity and model error and narrower distributions of climate sensitivity in the simple ensemble. These relationships, however, do not carry into the multi-model ensemble. This suggests that model weighting based on statistical relationships alone is unfounded, and perhaps that climate model errors are still large enough that model weighting is not sensible.

1. Model error and climate sensitivity

Equilibrium climate sensitivity, defined as the response in global-mean near-surface temperature to a doubling of atmospheric CO_2 concentrations from pre-industrial levels, is a useful proxy for climate change because many other projections scale with it. Climate models produce a range of estimates of climate sensitivity which can themselves be sensitive to fairly small changes in model formulation (Soden et al. 2004). The distribution of these projections has remained roughly the same for more than 30 years (compare, for example, Charney 1979; Solomon et al. 2007).

One might expect that with improvements of climate models over time, projections would converge to a narrower distribution, but this has not yet proved true: successive generations of climate models have produced improved simulations of the present-day climate (Reichler and Kim 2008) but commensurate distributions of climate sensitivity (Knutti et al. 2008).

The distribution might also be narrowed by invoking Bayes's theorem and weighting each prediction of climate sensitivity by the likelihood of the corresponding model (Murphy et al. 2004; Stainforth et al. 2005; Knutti et al. 2010). This likelihood is usually modeled as a decreasing function of model error, defined as some measure of the difference between long-term averages of observations and model simulations of the present-day climate. Weighting ensembles is fraught with theoretical issues including the impact of the sampling strategy used to construct the initial ensemble (Frame et al. 2005) and questions of how to treat an ensemble in which members have varying degrees of interdependence (e.g. Knutti et al. 2010; Tebaldi and Knutti 2007). But weighting projections has so far failed to substantially narrow distributions of climate sensitivity for a more practical reason: in current ensembles of climate models, global measures of error are not systematically related to climate sensitivity or the underlying feedbacks (Knutti et al. 2006;

Murphy et al. 2004; Piani et al. 2005; Sanderson et al. 2008; Collins et al. 2011).

Any observable measure of present-day error that is correlated with climate sensitivity in a given ensemble of climate projections, if used as a weight, would narrow the distribution of climate sensitivity estimates. This makes it tempting to seek such measures. But if the systematic relationships between the present day and the future in an ensemble of models have causes which are not shared by the physical climate system, weighting by such a measure can introduce substantial projection errors (Weigel et al. 2010).

Here we provide a practical demonstration of how hard it can be to determine whether relationships between the present day and the future in a given ensemble have a more general basis. We consider two ensembles of climate models: one containing a wide range of models and another employing a single model with varied values of closure parameters. We use the simpler, single-model ensemble as a proxy for understanding the behavior of the more complicated multi-model ensemble, much as one might use the more complicated ensemble to understand the real world. Section 2 describes the construction of the simple ensemble; we then show that this simple ensemble reproduces several relevant aspects of the multi-model ensemble. Section 4 describes the construction of a metric of present-day performance that is correlated with climate sensitivity in the simple model but does not generalize to the multi-model ensemble. We conclude by exploring the implications for model weighting.

2. A simple ensemble spanning a range of errors and climate sensitivities

We construct a perturbed-parameter ensemble by varying the values of selected closure parameters (Table 1) in physical parameterizations of the general circulation model ECHAM5 (Roeckner et al. 2003). The parameters are uncertain in observations and are those used to adjust the model so that its energy budget is balanced at the top of atmosphere (to within observational uncertainties and accounting for ocean heat storage). Each parameter is restricted to fairly small ranges near the default and all parameters are sampled simultaneously using Latin hypercube sampling (McKay et al. 1979). Five hundred realizations of ECHAM5 are created and each model is run for a single year using present-day climatological distributions of sea ice and sea surface temperature.

For each ensemble member, we compute an aggregate measure of the error in simulating the present-day distribution of clouds, radiation, and precipitation. Because it is not known which observable aspects, if any, of the present-day climate are connected to climate sensitivity, any aggregate metric is arbitrary; we justify the narrow focus of our choice by noting that a) differences in cloud feedbacks drive much of the diversity in climate sensitivity estimates from climate models (Soden and Held 2006), particularly by affecting the radiation budget, and b) a majority of the varied parameters are cloud-related. We compute the root-mean-square error relative to observations for cloud fraction, longwave and shortwave cloud radiative effects at the top of the atmosphere (e.g. Hartmann and Short 1980), and surface precipitation over each month of the annual cycle (Pincus et al. 2008). These errors are much larger in our short integrations than for long runs with well-tuned models because sampling errors are large. Still, the difference in errors based on individual years from longer runs (described below) is very small relative to the difference in error spanned by the ensemble, indicating that the diversity in error is robust. Errors in individual fields are standardized so that the distribution of each error across the ensemble has zero mean and a standard deviation of one, then added together to provide an aggregate error measure for each model, where low errors reflect greater skill relative to other members of the ensemble.

We sort the models according to this measure of aggregate error and compute the equilibrium climate sensitivity of every tenth model across the range of aggregate skill (so that the distribution of skill in the initial ensemble is roughly preserved). Ten-year runs are performed using a slab ocean model and present-day greenhouse gas concentrations, from which we determine the flux corrections necessary to maintain present-day sea surface temperatures. A fifty-year simulation is then performed using the same ocean heat flux corrections but with doubled carbon dioxide concentrations. Equilibrium climate sensitivity is computed as the difference in global mean surface temperature between the last ten years of the doubled CO_2 and the present-day simulations.

3. The simple ensemble as proxy for the multi-model ensemble

Results from this ensemble, in which all diversity arises from parametric uncertainty, are comparable in many ways to the multi-model ensemble from the World Climate Research Programme's Coupled Model Intercomparison Project phase 3 (CMIP3; see Meehl et al. 2007), which represents the majority of the world's climate models and contains both parametric and structural variability. In particular, the distributions of climate sensitivity (Figure 1a) and our aggregated measure of model error (Figure 1b) are similar in both ensembles. These quantities are not systematically related to each other in either ensemble (Figure 2). The similarity in the distributions of error and sensitivity, as well as the lack of a connection between the two, mirror previous experiences across a wide range of perturbed-parameter ensembles (Murphy et al. 2004; Stainforth et al. 2005; Collins et al. 2011).

The two ensembles also share an important structural feature: the same mechanism underlies the variability in climate sensitivity. In both ensembles, models with a large change in the net cloud radiative effect under doubled CO_2 concentrations are those with higher climate sensitivity (Figure 1a). The longwave cloud radiative effect in our ensemble does not change much between present-day and doubled CO_2 conditions, while diversity in shortwave cloud radiative effect (CRE_{SW}) changes, in turn, is largely driven by diversity in the response of low-latitude oceanic boundary layer clouds (Bony and Dufresne 2005).

By these measures, the perturbed-parameter ensemble is a successful proxy for the multi-model ensemble. This allows us to test the generality of model weighting techniques in two structurally distinct but statistically similar ensembles.

4. Developing measures of model error linked to climate sensitivity

We now design a measure of error in reproducing the present-day climate that is explicitly related to climate sensitivity in our simple ensemble. We identify such a measure by focussing on the low-latitude oceanic boundary layer clouds whose response is tightly linked to climate sensitivity (Bony and Dufresne 2005). Boundary layer clouds dominate CRE_{SW} in subsidence regions, i.e. where the mid-tropospheric pressure velocity is downward ($\omega_{500} > 0$), so we sort present-day CRE_{SW} by this quantity (Bony et al. 2004). In our ensemble the present-day distribution of

CRE_{SW} in subsidence regions differs markedly between the ten highest- and ten lowest-sensitivity model variants (Figure 3a). Higher sensitivity models have weaker values of CRE_{SW}, indicating that clouds are some combination of less frequent, less extensive, or less reflective than in low-sensitivity simulations. The higher sensitivity models are also more consistent with observations (here, cloud radiative effect derived from satellite observations (Wielicki et al. 1996; Loeb et al. 2009) and sorted by ω_{500} inferred from ERA-Interim reanalysis data (Simmons et al. 2007)). Although the highest- and lowest-sensitivity models in our ensemble are distinct from each other, at the most frequent values of subsidence essentially all members over-estimate CRE_{SW} relative to observations. In regions of large-scale ascent ($\omega_{500} < 0$) the distributions of CRE_{SW} in the highest- and lowest-sensitivity models are much broader and overlap significantly.

In nature, boundary layer clouds in subsiding regions over the oceans are further correlated (Medeiros and Stevens 2011) with lower tropospheric thermodynamic stability (LTS; see Bretherton and Wyant 1997; Klein and Hartmann 1993), here defined as the difference in the potential temperature at 1000 hPa and 700 hPa. Our simple ensemble reproduces this dependency as well (Figure 3b). Through much of the range of LTS the highest- and lowest-sensitivity models are indistinguishable, but in the range 13 K $<$ LTS $<$ 17 K CRE_{SW} in the high-sensitivity models is consistently weaker, and in better agreement with observations, than for low-sensitivity models. These are the most frequent values of LTS in subsiding regions in our ensemble.

Figure 3 demonstrates why global measures of skill are unrelated to model climate sensitivity: because the clouds whose systematic changes explain the diversity in sensitivity occur in a small region of the globe. Most measures of skill compare models to observations in global domains (e.g. Gleckler et al. 2008; Pincus et al. 2008; Reichler and Kim 2008). Restricting the geographical domain over which errors are computed would not change this result much: even considering only the low-latitude oceans, the root-mean-square difference with observations are influenced not only by the regions controlling the sensitivity but also by ascending regions, where errors are large, and low-sensitivity models perform somewhat better, on average.

We define instead a conditioned error measure E_c as the root-mean-square difference between model simulations and observations of CRE$_{SW}$ integrated over regions with large-scale subsidence ($\omega_{500} > 0.03$ Pa s^{-1}) and moderate lower tropospheric stability (13 K $<$ LTS $<$ 17 K). Regions satisfying both conditions comprise just 5% of the area of the tropics (2.5% of the globe) in the observations and somewhat more in the models. Nonetheless, E_c is a reasonably good predictor of climate sensitivity in the simple ensemble (Figure 4), which means it can be used to narrow the distribution of climate sensitivity estimates. Figure 4b shows the distribution of climate sensitivity obtained from the perturbed-parameter ensemble before and after weighting by the likelihood $L(E_c) = \exp(-E_c/2)$ (Murphy et al. 2004). The standard deviation of the posterior distribution is 3/4 of that of the prior distribution, mostly because a few models with low sensitivity have large errors and hence low weight. The mean climate sensitivity also increases by 0.35 K.

But despite the many similarities between the perturbed-parameter and multi-model ensembles, the systematic relationship between climate sensitivity and E_c does not carry into the multi-model ensemble (Figure 5), nor does the distribution of sensitivity estimates from the multi-model ensemble change when weighted by $L(E_c)$.

5. Implications for weighting projections from multi-model ensembles

One could conclude that we have obtained a null result and that the single-model perturbed-parameter ensemble is, after all, a poor proxy for the multi-model ensemble. Instead, we propose that these calculations are a concrete illustration of some of the issues involved in the weighting and more general interpretation of multi-model ensembles.

First, our results confirm that it is possible to obtain distributions of climate sensitivity and global measures of error as diverse as those produced by the multi-model ensemble with even modest variations about a single model. This suggests that variability in error and sensitivity at these levels is easy to come by (though why this is so remains an intriguing open question). In fact, in our ensemble diversity in skill and climate sensitivity arises from surprisingly simple parametric sensitivity: Climate sensitivity is primarily related to the entrainment rate for shallow convection, which varies along with a cloud mass flux parameter (explaining 44% of the variance in climate sensitivity; Table 1), while aggregate error is related to another parameter, the entrainment rate for deep convection (explaining 64% of the variance in aggregated error; Table 1). If broad diversity in behavior can arise from underlying simplicity then the diversity itself is uninformative. This is an illustrative reminder that the distribution of climate sensitivity from any model ensemble can not be interpreted as an estimate of the total uncertainty in climate sensitivity.

Second, while the motivation to narrow the distribution of climate sensitivity estimates is strong, our results dramatize the danger of focusing exclusively on this goal. Relationships between sensitivity and model fidelity in any ensemble emerge from an unknown mix of underlying similarity in model representation and error, statistical sampling error, and physical relationships also present in the natural world. This means that arbitrarily-chosen error measures may arise from underlying similarity not present in the physical climate system. We argue that because metrics developed from the full multi-model ensemble alone can not be falsified by comparison to more general ensembles, they can not be justified as a model likelihood purely on the basis of the strength of the statistical connection between that metric and climate sensitivity. Indeed, where observations have been used successfully to constrain model response (Hall and Qu 2006; Clement et al. 2009) statistical metrics have been bolstered by physical arguments. Much depends on the way weights are chosen, since incorrect weighting (that is, weighting not related to true model likelihood) can substantially reduce the benefits of using an ensemble of projections (Weigel et al. 2010).

Finally, it is possible that present-day models are not yet sufficiently accurate to benefit from model weighting. Weighting model projections by skill is an assertion that models are likely to produce accurate estimates of future climate in proportion to their ability to reproduce some aspects of the present-day climate; the implicit assumption is that models with higher skill are more likely to be accurate representations of the physical climate system. But by most measures, no current climate model produces distributions of the present-day climate statistically consistent with observations (Gleckler et al. 2008; Pincus et al. 2008, see also Figure 3 and 5), implying that all models are formally unlikely. Weighting an ensemble under these circumstances is essentially asserting that incorrect models are more reliable than even-more-incorrect models. But the result of Bayes's theorem is ambiguous when the system being modeled is far from the system being observed, and it may be that model weighting will be more profitable when the collection of models we have is closer to the world we observe.

Acknowledgments.

We thank the Max Planck Society, the International Max Planck Research School for Earth System Modeling, the National Science Foundation's Center for Multi-scale Modeling of Atmospheric Phenomena, and the German Research Foundation Emmy Noether grant program for supporting this work. Bjorn Stevens, Louise Nuijens, Thorsten Mauritsen, and Jeffrey L. Anderson provided valuable feedback on early drafts of this paper and two anonymous reviewers helped us refine the arguments. We acknowledge the modeling groups, PCMDI and the WCRP's Working Group on Coupled Modeling (WGCM) for their roles in making the WCRP CMIP3 multi-model dataset available. The Office of Science, U.S. Department of Energy, provides support for the CMIP3 dataset. Model simulations were carried out on the supercomputing facilities of the German Climate Computation Centre (DKRZ) in Hamburg.

REFERENCES

Bony, S. and J.-L. Dufresne, 2005: Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models. *Geophys. Res. Lett.*, **32 (20)**, L20 806, doi:10.1029/2005gl023851.

Bony, S., J.-L. Dufresne, H. L. Treut, J.-J. Morcrette, and C. Senior, 2004: On dynamic and thermodynamic components of cloud changes. *Climate Dynamics*, **22 (2-3)**, 71–86, doi:10.1007/s00382-003-0369-6.

Bretherton, C. S. and M. C. Wyant, 1997: Moisture transport, lower-tropospheric stability, and decoupling of cloud-topped boundary layers. *Journal of the Atmospheric Sciences*, **54 (1)**, 148–167, doi:10.1175/1520-0469(1997)054⟨0148:MTLTSA⟩2.0.CO;2.

Cahalan, R. F., W. Ridgway, W. J. Wiscombe, T. L. Bell, and J. B. Snider, 1994: The albedo of fractal stratocumulus clouds. *Journal of the Atmospheric Sciences*, **51 (16)**, 2434–2455, doi:10.1175/1520-0469(1994)051⟨2434:TAOFSC⟩2.0.CO;2.

Charney, J., 1979: Carbon dioxide and climate: A scientific assessment. *National Acadamy of Sciences*.

Clement, A. C., R. Burgman, and J. R. Norris, 2009: Observational and model evidence for positive low-level cloud feedback. *Science*, **325 (5939)**, 460–464, doi:10.1126/science.1171255.

Collins, M., B. Booth, B. Bhaskaran, G. Harris, J. Murphy, D. Sexton, and M. Webb, 2011: Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model ensembles. *Climate Dynamics*, to appear, doi:10.1007/s00382-010-0808-0.

Frame, D. J., B. B. B. Booth, J. A. Kettleborough, D. A. Stainforth, J. M. Gregory, M. Collins, and M. R. Allen, 2005: Constraining climate forecasts: The role of prior assumptions. *Geophysical Research Letters*, **32**, L09 702, doi:10.1029/2004GL022241.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113 (D6)**, D06 104, doi:10.1029/2007jd008972.

Hall, A. and X. Qu, 2006: Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophysical Research Letters*, **33**, doi:10.1029/2005GL025127.

Hartmann, D. L. and D. A. Short, 1980: On the use of earth radiation budget statistics for studies of clouds and climate. *Journal of the Atmospheric Sciences*, **37 (6)**, 1233–1250, doi:10.1175/1520-0469(1980)037⟨1233:OTUOER⟩2.0.CO;2.

Klein, S. A. and D. L. Hartmann, 1993: The seasonal cycle of low stratiform clouds. *Journal of Climate*, **6 (8)**, 1587–1606, doi:10.1175/1520-0442(1993)006⟨1587:TSCOLS⟩2.0.CO;2.

Knutti, R., R. Furrer, C. Tebaldi, J. Cermak, and G. A. Meehl, 2010: Challenges in combining projections from multiple climate models. *Journal of Climate*, **23 (10)**, 2739–2758, doi:10.1175/2009JCLI3361.1.

Knutti, R., G. A. Meehl, M. R. Allen, and D. A. Stainforth, 2006: Constraining climate sensitivity from the seasonal cycle in surface temperature. *Journal of Climate*, **19 (17)**, 4224–4233, doi:10.1175/2007jcli2119.1.

Knutti, R., et al., 2008: A review of uncertainties in global temperature projections over the twenty-first century. *Journal of Climate*, **21 (11)**, 2651–2663, doi:10.1175/2009JCLI3361.1.

Loeb, N. G., B. A. Wielicki, D. R. Doelling, G. L. Smith, D. F. Keyes, S. Kato, N. Manalo-Smith, and T. Wong, 2009: Toward optimal closure of the earth’s top-of-atmosphere radiation budget. *Journal of Climate*, **22 (3)**, 748–766, doi:10.1175/2008JCLI2637.1.

Lott, F., 1999: Alleviation of stationary biases in a GCM through a mountain drag parameterization scheme and a simple representation of mountain lift forces. *Monthly Weather Review*, **127 (5)**, 788–801, doi:10.1175/1520-0493(1999)127⟨0788:AOSBIA⟩2.0.CO;2.

McKay, M. D., R. J. Beckman, and W. J. Conover, 1979: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21 (2)**, 239–245, URL `http://www.jstor.org/stable/1268522`.

Medeiros, B. and B. Stevens, 2011: Revealing differences in GCM representations of low clouds. *Climate Dynamics*, **36**, 385–399, doi:10.1007/s00382-009-0694-5.

Meehl, G. A., C. Covey, K. E. Taylor, T. Delworth, R. J. Stouffer, M. Latif, B. McAvaney, and J. F. B. Mitchell, 2007: The WCRP CMIP3 multimodel dataset: A new era in climate change research. *Bulletin of the American Meteorological Society*, **88 (9)**, 1383–1394, doi:10.1175/BAMS-88-9-1383.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430 (7001)**, 768–772, doi:10.1038/nature02771.

Piani, C., D. J. Frame, D. A. Stainforth, and M. R. Allen, 2005: Constraints on climate change from a multi-thousand member ensemble of simulations. *Geophys. Res. Lett.*, **32 (23)**, L23 825, doi:10.1029/2005gl024452.

Pincus, R., C. P. Batstone, R. J. P. Hofmann, K. E. Taylor, and P. J. Glecker, 2008: Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. *J. Geophys. Res.*, **113 (D14)**, D14 209, doi:10.1029/2007jd009334.

Reichler, T. and J. Kim, 2008: How well do coupled models simulate today's climate? *Bulletin of the American Meteorological Society*, **89 (3)**, 303–311, doi:10.1175/BAMS-89-3-303.

Roeckner, E., et al., 2003: The atmospheric general circulation model ECHAM5. Part I: Model description. Tech. Rep. 349, Max-Planck-Institut für Meteorologie, Hamburg, Germany.

Sanderson, B., C. Piani, W. Ingram, D. Stone, and M. Allen, 2008: Towards constraining climate sensitivity by linear analysis of feedback patterns in thousands of perturbed-physics GCM simulations. *Climate Dynamics*, **30 (2)**, 175–190, doi:10.1007/s00382-007-0280-7.

Simmons, A., S. Uppala, D. Dee, and K. S., 2007: ERA-Interim: New ECMWF reanalysis products from 1989 onwards: New ECMWF reanalysis products from 1989 onwards. *ECMWF Newsletter*, **110**, 29–35.

Soden, B. J., A. J. Broccoli, and R. S. Hemler, 2004: On the use of cloud forcing to estimate cloud feedback. *Journal of Climate*, **17**, 3661–3665, doi:10.1175/1520-0442(2004)017⟨3661:OTUOCF⟩2.0.CO;2.

Soden, B. J. and I. M. Held, 2006: An assessment of climate feedbacks in coupled ocean-atmosphere models. *Journal of Climate*, **19**, 3354–3360, doi:10.1175/JCLI3799.1.

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K. Averyt, M. Tignor, and H. Miller, 2007: Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge University Press*.

Stainforth, D. A., et al., 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, **433 (7024)**, 403–406, doi:10.1038/nature03301.

Stephens, G. L., S.-C. Tsay, P. W. Stackhouse, and P. J. Flatau, 1990: The relevance of the microphysical and radiative properties of cirrus clouds to climate and climatic feedback. *Journal of the Atmospheric Sciences*, **47 (14)**, 1742–1754, doi:10.1175/1520-0469(1990)047⟨1742:TROTMA⟩2.0.CO;2.

Tebaldi, C. and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A*, **365**, 2053–2075, doi:10.1098/rsta.2007.2076.

Tiedtke, M., 1989: A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Monthly Weather Review*, **117 (8)**, 1779–1800, doi:10.1175/1520-0493(1989)117⟨1779:ACMFSF⟩2.0.CO;2.

Weigel, A. P., R. Knutti, M. A. Liniger, and C. Appenzeller, 2010: Risks of model weighting in multimodel climate projections. *Journal of Climate*, **23 (15)**, 4175–4191, doi:10.1175/2010JCLI3594.1.

Wielicki, B. A., B. R. Barkstrom, E. F. Harrison, R. B. Lee, G. Louis Smith, and J. E. Cooper, 1996: Clouds and the Earth's Radiant Energy System (CERES): An Earth Observing System

Experiment. *Bulletin of the American Meteorological Society*, **77 (5)**, 853–868, doi:10.1175/1520-0477(1996)077⟨0853:CATERE⟩2.0.CO;2.

List of Tables

1 List of perturbed parameters in the ECHAM5 ensemble, their description, default value, the range they are varied in and the percentage contribution to the variation in skill and climate sensitivity. [1]Default value in the atmosphere-only model. [2] Default value in the coupled model. *Indicates coupled parameters, to keep top of the atmosphere radiative fluxes close to balance. 18

TABLE 1. List of perturbed parameters in the ECHAM5 ensemble, their description, default value, the range they are varied in and the percentage contribution to the variation in skill and climate sensitivity. [1]Default value in the atmosphere-only model. [2] Default value in the coupled model. *Indicates coupled parameters, to keep top of the atmosphere radiative fluxes close to balance.

Description of parameter	*Default value*	*Range*	R^2[%] *Skill*	R^2[%] *Sensitivity*
Entrainment rate for shallow convection* (Tiedtke 1989)	0.0003	0.0003 - 0.001	3	44
Cloud mass flux above level of non-buoyanc* (Tiedtke 1989)	$0.1^1/0.3^2$	0.1 - 0.3333	3	44
Entrainment rate for penetrative convection (Tiedtke 1989)	0.0001	0.00001 - 0.0005	64	0
Conversion rate from cloud water to rain (Tiedtke 1989)	0.0004	0.0001-0.005	0	1
In-homogeneity of liquid clouds (Cahalan et al. 1994)	0.7	0.65 - 1	4	0
In-homogeneity of ice clouds (Cahalan et al. 1994)	$0.7^1/0.8^2$	0.65 - 1	20	1
Asymmetry of ice particles in clouds (Stephens et al. 1990)	$0.91^1/0.85^2$	0.75 - 1	0	1
Coefficient for horizontal diffusion	12	6 - 24	6	5
Gravity wave drag activation threshold (mean) (Lott 1999)	500	400 - 1000	2	0
Gravity wave drag activation threshold (stddev) (Lott 1999)	200	100 - 700	2	0
Albedo minimum of snow/ice	0.6/0.5	0.45 - 0.65	8	0
Albedo maximum of snow/ice	0.8/0.75	0.75 - 0.9	9	3

List of Figures

1 Climate sensitivity and skill in two ensembles of climate models. a) Equilibrium climate sensitivity as a function of the change in global annual mean net cloud radiative effect (ΔCRE) under doubled CO_2 conditions. The CMIP3 ensemble is shown with red dots; the models are also labelled. The distribution of climate sensitivities is similar in the two ensembles, as is the mechanism driving the variability (the change in cloud radiative effect). Background colors indicate the highest (red) and lowest (blue) sensitivity models used later. b) Distributions of aggregate skill in present-day simulations of clouds, radiation, and precipitation for our perturbed-parameter ensemble (histogram) and from the CMIP3 ensemble (dots). The skill measure integrates over the annual cycle, the geographic distribution, and four variables. Black dots indicate the performance of the base ECHAM model (atmosphere-only and coupled to an ocean model) within the CMIP3 ensemble. 22

2 Global measure of skill, aggregated over cloud radiative effects, precipitation and cloud cover are unrelated to climate sensitivity in a simple ensemble and the multi-model CMIP3 ensemble. 23

3 Relationships between present-day cloud properties and atmospheric state in a perturbed-parameter ensemble. Both figures are restricted to the tropical (30°S – 30°N) oceans. The ten highest- and lowest-sensitivity models (red and blue, respectively) in the perturbed-parameter ensemble are shown; box and whisker plots summarize the medians (central lines), quartiles (box ends), and range (whiskers) of the distributions. Observations are shown in black, and the frequency distribution of models and observations in the lower part of each panel. a) Monthly-mean values of shortwave cloud radiative effect CRE$_{SW}$ (all-sky fluxes minus clear-sky fluxes) sorted by mid-tropospheric pressure velocity ω_{500}. Boundary-layer clouds dominate in subsiding ($\omega_{500} > 0$) regions where high- and low-sensitivity models in our ensemble are distinct. Global measures of skill, though, are dominated by the errors unrelated to climate sensitivity occurring through the entire domain. The grey area indicates regions used in figure 3b. b) Cloud radiative effect in subsidence regions ($\omega_{500} > 0.03$ Pa s^{-1}) sorted by lower tropospheric stability. The grey background color indicates regions used for weighting in figure 4b. High- and low-sensitivity models are distinct through a 4 K range of stability, though the ensemble is systematically roughly 2 K less stable than is observed. 24

4 A tightly-focused measure of skill narrows the distribution of climate sensitivity in a simple ensemble. a) Equilibrium climate sensitivity as a function of conditionally-sampled root-mean-square error in shortwave cloud radiative effect of simulations compared to satellite observations. The error is computed only in regions of descending air ($\omega_{500} > 0.03$ Pa s^{-1}) and moderate lower tropospheric thermodynamic stability (13 K $<$ LTS $<$ 17 K) over tropical oceans. b) Distributions of climate sensitivity estimates before (black) and after weighting by a function of the error in panel a. Weighting by this metric decreases the standard deviation of the distribution by about 23% and increases the mean by 0.35 K. 25

5 Relationships between present-day cloud properties and atmospheric state in a multi-model ensemble. These figures are constructed in the same way as figure 3, but the distribution of cloud radiative effect as sorted by ω_{500} (a) or lower tropospheric stability in subsiding regions (b) does not distinguish between high- and low-sensitivity models in the CMIP3 ensemble. 26

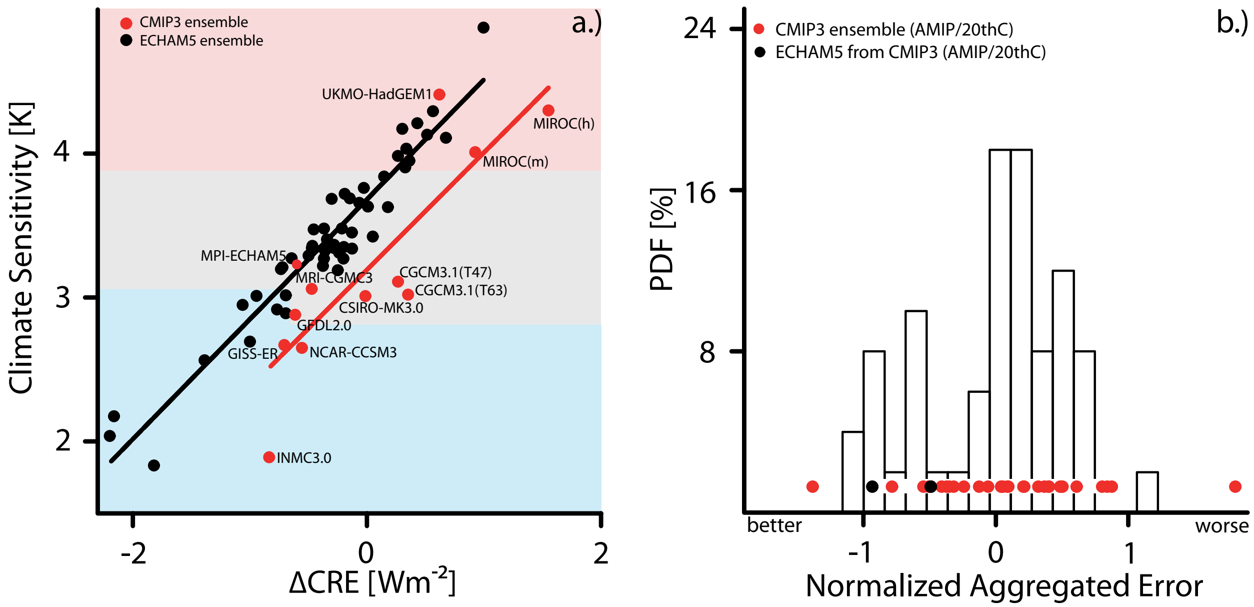

FIG. 1. Climate sensitivity and skill in two ensembles of climate models. a) Equilibrium climate sensitivity as a function of the change in global annual mean net cloud radiative effect (ΔCRE) under doubled CO_2 conditions. The CMIP3 ensemble is shown with red dots; the models are also labelled. The distribution of climate sensitivities is similar in the two ensembles, as is the mechanism driving the variability (the change in cloud radiative effect). Background colors indicate the highest (red) and lowest (blue) sensitivity models used later. b) Distributions of aggregate skill in present-day simulations of clouds, radiation, and precipitation for our perturbed-parameter ensemble (histogram) and from the CMIP3 ensemble (dots). The skill measure integrates over the annual cycle, the geographic distribution, and four variables. Black dots indicate the performance of the base ECHAM model (atmosphere-only and coupled to an ocean model) within the CMIP3 ensemble.

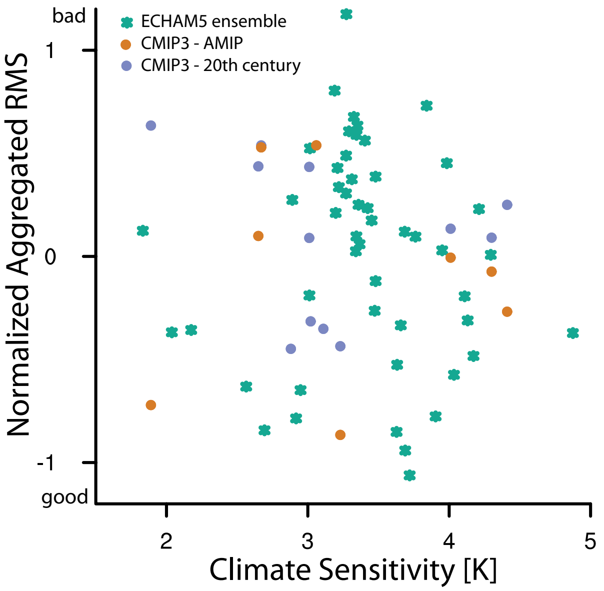

FIG. 2. Global measure of skill, aggregated over cloud radiative effects, precipitation and cloud cover are unrelated to climate sensitivity in a simple ensemble and the multi-model CMIP3 ensemble.

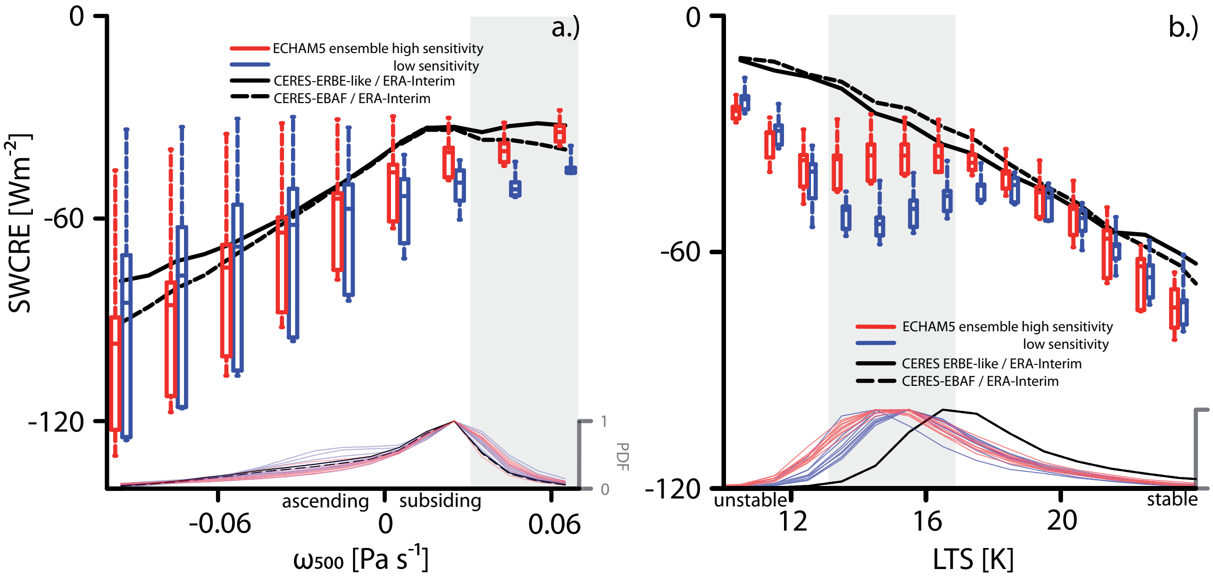

FIG. 3. Relationships between present-day cloud properties and atmospheric state in a perturbed-parameter ensemble. Both figures are restricted to the tropical (30°S – 30°N) oceans. The ten highest- and lowest-sensitivity models (red and blue, respectively) in the perturbed-parameter ensemble are shown; box and whisker plots summarize the medians (central lines), quartiles (box ends), and range (whiskers) of the distributions. Observations are shown in black, and the frequency distribution of models and observations in the lower part of each panel. a) Monthly-mean values of shortwave cloud radiative effect CRE$_{SW}$ (all-sky fluxes minus clear-sky fluxes) sorted by mid-tropospheric pressure velocity ω_{500}. Boundary-layer clouds dominate in subsiding ($\omega_{500} > 0$) regions where high- and low-sensitivity models in our ensemble are distinct. Global measures of skill, though, are dominated by the errors unrelated to climate sensitivity occurring through the entire domain. The grey area indicates regions used in figure 3b. b) Cloud radiative effect in subsidence regions ($\omega_{500} > 0.03$ Pa s^{-1}) sorted by lower tropospheric stability. The grey background color indicates regions used for weighting in figure 4b. High- and low-sensitivity models are distinct through a 4 K range of stability, though the ensemble is systematically roughly 2 K less stable than is observed.

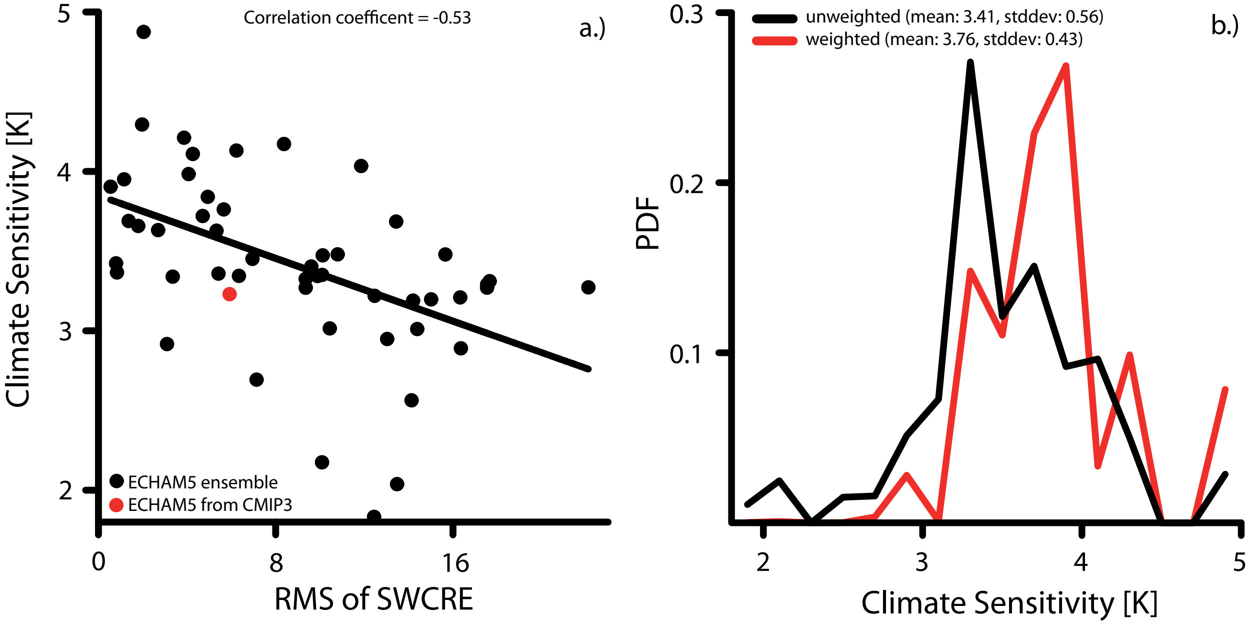

FIG. 4. A tightly-focused measure of skill narrows the distribution of climate sensitivity in a simple ensemble. a) Equilibrium climate sensitivity as a function of conditionally-sampled root-mean-square error in shortwave cloud radiative effect of simulations compared to satellite observations. The error is computed only in regions of descending air ($\omega_{500} > 0.03$ Pa s^{-1}) and moderate lower tropospheric thermodynamic stability (13 K $<$ LTS $<$ 17 K) over tropical oceans. b) Distributions of climate sensitivity estimates before (black) and after weighting by a function of the error in panel a. Weighting by this metric decreases the standard deviation of the distribution by about 23% and increases the mean by 0.35 K.

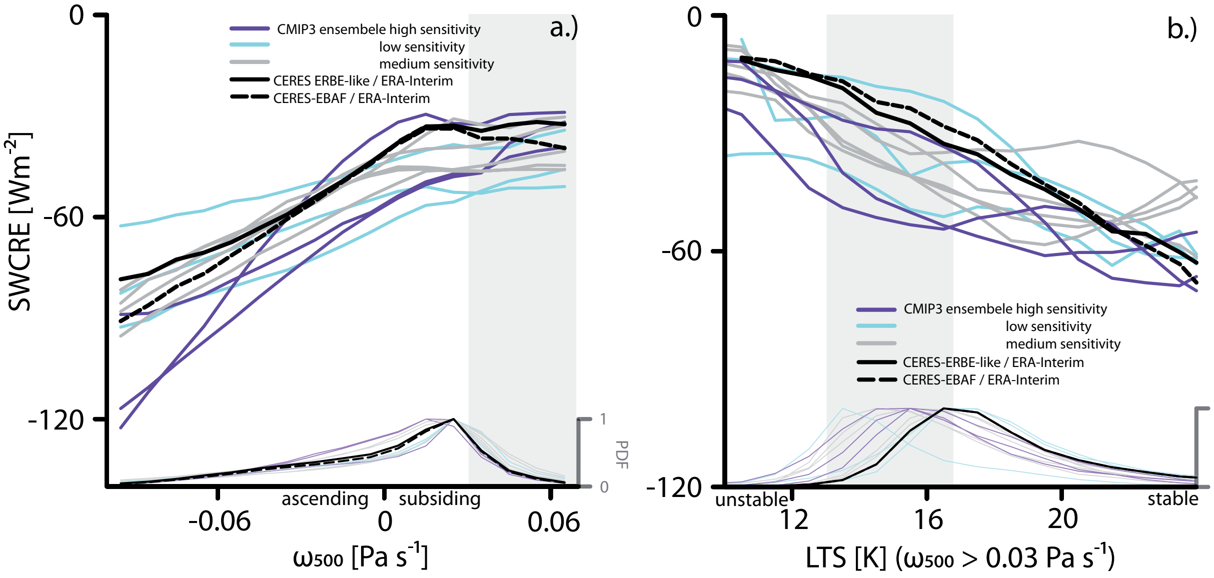

FIG. 5. Relationships between present-day cloud properties and atmospheric state in a multi-model ensemble. These figures are constructed in the same way as figure 3, but the distribution of cloud radiative effect as sorted by ω_{500} (a) or lower tropospheric stability in subsiding regions (b) does not distinguish between high- and low-sensitivity models in the CMIP3 ensemble.